# Finding causal variants for complex genetic disease

Spijker, Geert Theodoor

3

# Hidden first-degree relatedness between subjects in genetic population studies can be detected using routine genotyping data

Geert T. Spijker[1], Gerjan Navis[2] and Gerard J. te Meerman[1]
Departments of [1]Medical Genetics and [2]Nephrology, University Medical Centre Groningen, the Netherlands

*To find genetic associations between markers and disease in complex genetic disease, large samples of cases and controls need to be compared and many genotypes need to be determined. Hidden relatedness within such samples could bias comparisons based on assumptions of random sampling. We propose a statistical methodology to establish first-degree relatedness within samples and to estimate the numbers of markers needed to do so at a desired level of confidence, also in the presence of missing data and genotyping errors.*

*The approach is illustrated with a data set from the PREVEND population study conducted in the city of Groningen, the Netherlands. In a random sample (N= 638) four related subject pairs were found. However, in a sub sample that was non-genetically enriched for containing relatives (N= 590), 14 related subject pairs were found; a statistically significant ($p<0.05$) increase. Results suggest that genetic case-control studies might indeed contain relevant numbers of relatives. However, routine genotyping data can be used to detect hidden relatedness.*

## Introduction

Population-based genetic studies use statistics that suppose the subjects to be unrelated. P-values are computed, assuming affected subjects to be just as distantly related to each other as unaffected subjects. This assumption is violated, in the case of a well-known potential bias in case-control association studies: population stratification. When sub-population A has a higher incidence of disease than the rest of the population, a case-control design will preferentially select cases from A, increasing average relatedness among cases. However, the assumption can very well be violated in the absence of population stratification. For example, when recruiting the participants of a case-control study from a limited population, a certain number of (closely) related individuals will be included. When recruitment for the study is voluntary, the number of relatives might be further increased, as willingness to participate in a study could show a familial component. In genetic case-control

studies, subjects with the trait are preferably sampled, while the trait or disease under study supposedly has a hereditary component. This could lead to excess numbers of relatives, especially among the affected subjects [De Visser CL, Te Meerman GJ, Meyboom-de Jong B, De Visser W, Bilo HJG, submitted]. Relatedness will lead to increased genetic concordance between subjects and increased deviations from expected allele frequencies, also at loci not involved in the trait under study. Ignoring relatedness between subjects that exceeds random expectation could thus increase the number of false-positive results; the increased genetic concordance between hidden relatives could be mistaken for a causal genetic effect. These effects are most pronounced for first-degree relatives. Therefore, it is of interest to be able to identify hidden first-degree relatedness within a genetic case-control study. Additionally, knowledge of relatedness between subjects could improve assignment of gametic phase, especially when Expectation-Maximization (EM)-algorithms are used. This might be useful when performing genetic multilocus studies.

In this paper we propose a strategy to detect related subjects in a sample, and subsequently apply this strategy to estimate the frequency of first-degree relatedness in a genetic association study. The sample consists of 1228 subjects from the population of the middle-sized town of Groningen, the Netherlands (180 604 inhabitants in 2005 [1]. Attention will be given to the influence of missing data and genotyping errors.

## Theoretical considerations of genetically detecting parent-offspring pairs in a sample of subjects.

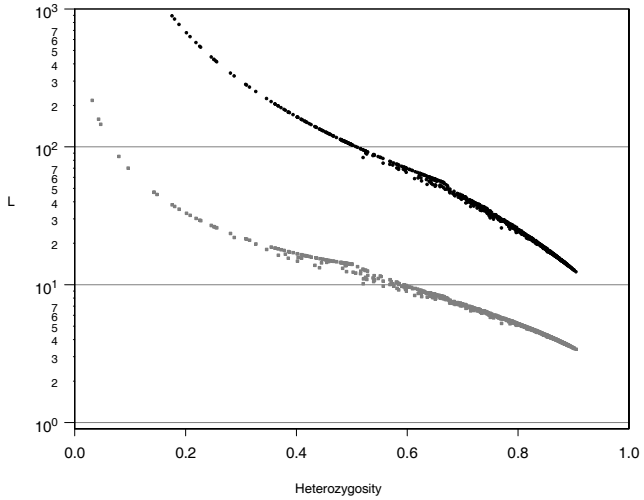
*Number of subject pairs*

The analysis starts with a sample of N subjects. All N· (N-1)/2 possible pairs of subjects have to be considered in turn, in order to identify pairs that are related. The number of pairs to be considered is proportional to the square of the number of subjects; it rises quickly with increasing sample size. In face of this large number of pairs, a test identifying related subject-pairs should be very specific to minimize false-positive results.

*Identifying monozygotic twins*

Monozygotic (MZ) twins are genetically identical, and are therefore easy to detect: in the absence of somatic marker mutations, they share all alleles at all loci. In practice, marker mutations are quite rare (<1/1000 per meiosis [2]), but genotyping errors occur more often and cannot be ignored. When marker mutations can be assumed to be absent, known MZ twins can be used to estimate an overall genotyping error rate:

**Figure 1: Theoretical number of markers needed**




On the y-axis ([10]log-scale) the number of independent markers *L* needed to reach a false-positive rate $\alpha$ of $10^{-6}$ against the heterozygosity of the markers on the x-axis. Shown are the results for one thousand simulated random allele frequency distributions, ignoring genotyping errors. Parent-offspring pairs are shown in black, monozygotic twins in grav

the number of differences between the genotypes of both subjects, divided by twice the number of informative loci (as two genotypes are produced for each informative locus).

*Identifying parent-offspring pairs*

On each genetic locus, offspring inherit one allele from each parent. Two persons that are related as a parent-offspring pair (PO pair) will therefore have at least one allele identical at each marker. We will call loci that meet this criterion: consistent loci. At co-dominant markers, such as microsatellite markers, both alleles in a heterozygote can be identified. Thus, when a pair of subjects has a locus *without* identical alleles, it cannot form a PO pair – at least, provided that no genotyping error or marker mutation has occurred. When allowing for genotyping errors, subject pairs with a small number of inconsistent loci could in reality still be PO pairs. Loci with one or more unknown alleles are not informative.

The genetic relation between siblings is not deterministic. On each locus sibs will share 0, 1 or 2 parental alleles with probability 0.25, 0.5 and 0.25, respectively. As markers are not completely informative with respect to parental status, the actual allele sharing will be higher. Direct identification of sibs should rely on an increased average allele sharing (with an appropriate standard deviation) in comparison to the rest of the sample. We will not further pursue this issue. The fact that sibs share their

**Table 1: Theoretical number of markers needed**

| Heterozygosity | MZ twins | | PO pair | |
|---|---|---|---|---|
| | $p_{sh}$ | *N markers* | $p_{sh}$ | *N markers* |
| 0.50 | 0.353 | 13.3 | 0.872 | 100.8 |
| 0.60 | 0.229 | 9.4 | 0.813 | 67.0 |
| 0.70 | 0.145 | 7.2 | 0.726 | 43.2 |
| 0.80 | 0.069 | 5.2 | 0.573 | 24.8 |
| 0.90 | 0.020 | 3.6 | 0.353 | 13.3 |

The chance $p_{sh}$ of two random subjects sharing both alleles (MZ twins), or at least one allele (PO pair) was calculated for simulated markers with random allele frequencies. Average values were taken for 20 markers with heterozygosity near the stated value. Computed is the number of markers needed to reach $\alpha = 10^{-6}$.

**Table 2: Expected numbers of inconsistent loci depending on error rate**

| Error rate | Number of inconsistent loci | | | |
|---|---|---|---|---|
| | *0* | *1* | *2* | *More* |
| 0.001 | 0.942 | 0.057 | 0.002 | 0.000 |
| 0.005 | 0.740 | 0.223 | 0.033 | 0.003 |
| 0.010 | 0.547 | 0.332 | 0.099 | 0.022 |

Assuming a binomial distribution and $L$=30, the fractions of PO pairs showing 0, 1 and 2 inconsistent loci are shown, using the frequency of detectable genotyping errors in the first column.

parents enables indirect identification of sibs, on condition that at least one parent is included in the sample.

*Number of markers*

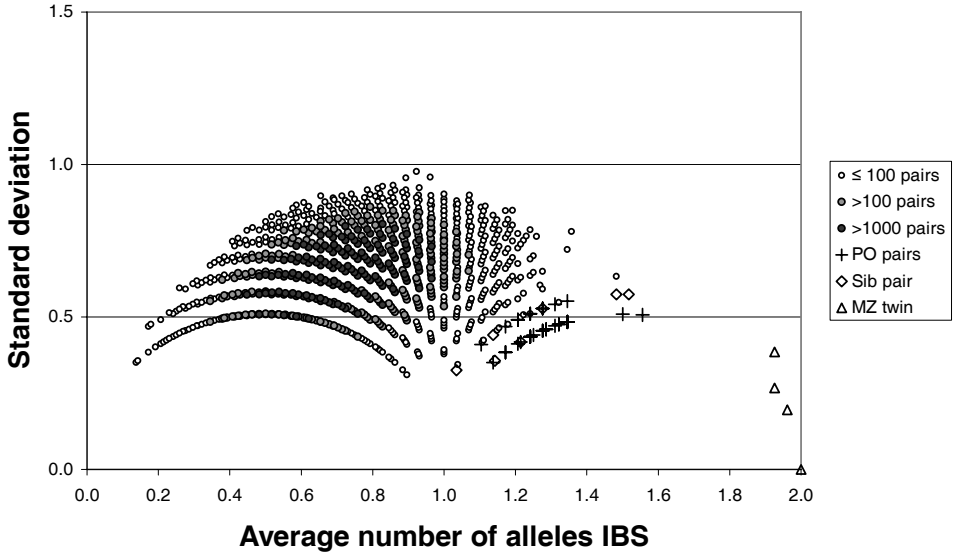
How many markers are needed to reliably distinguish between related and random subject pairs? When the allele frequency distributions of the markers are known, this value can analytically be computed, see **appendix 1**. Figure 1 and table 1 illustrate the resulting numbers.

Of course, these estimations have limited value. The assumption of identical allele frequencies for all markers leads to an overestimation; when the markers vary in heterozygosity, the markers with high heterozygosity have relatively great impact, reducing the actual number of markers. It is numerically worked out in **appendix 1**, but a simple example can help: a set of uninformative markers can become completely informative by adding one marker that is completely informative.

*Genotyping errors*

When a real life data set is to be analyzed, genotyping errors cannot be ignored. In MZ twin pairs, each genotyping error will be detectable (ignoring the possibility that exactly the same error is made in both individuals). For PO pairs, at most half of the

**Figure 2: Amount of alleles IBS between subject pairs**




Output of GRR [8] slightly modified. For all subject pairs the average number of alleles identical-by-state (IBS) (x-axis) and the matching standard deviation (SD, y-axis) is computed. All pairs with more than 25 informative loci are shown. When more than 100 subject pairs are observed with a specific combination of alleles IBS and SD, circles are filled gray. When more than 1000 subject pairs are observed circles are filled black. MZ twin pairs are positioned in the lower right corner (triangles). Plus sign: probably PO pairs (based on the analyses described in the text). Open diamonds: probably sib pairs (based on the analyses described in the text).

errors will lead to an inconsistent locus [3, 4], as a parent transmits only one allele to each child, and offspring receive only one allele from each parent, and markers are not completely informative. What fraction of genotyping errors will be exactly detectable depends on both the allele frequency distribution and the type of errors made. The rate at which detectable genotyping errors have been made can be estimated from the data.

Assuming a binomial distribution for genotyping errors, it can be computed what fraction of PO pairs are expected with 0, 1, 2 and more inconsistent loci. Let $\varepsilon$ be the rate per produced genotype at which detectable genotyping errors are made, then $(1-\varepsilon)^{2L}$ is the rate at which no error is made, while $2L*\varepsilon*(1-\varepsilon)^{2L-1}$ is the rate at which exactly 1 locus is consistent, etc. These fractions are shown in **Table 2** for $L$=30. At realistic genotyping error rates, the great majority of the pairs have at most 2 inconsistent loci.

The influence of genotyping errors on random subject pairs and their chance to appear related is much less profound, and depends on the type of error made. Let us consider three models:

1) The real genotype is scored as a random other genotype, with, in most cases, only one erroneous allele. In that case, genotyping errors do not affect the probability of false-positive results, $\alpha$.
2) Genotyping errors produce non-random alleles, e.g. when non variable by-products are scored as allele. When that by-product does not match a frequent allele, errors increase (appearing) marker heterozygosity and reduce the probability of false positive results, $\alpha$, unless errors are very common.
3) However, the erroneous allele might be scored as an existing allele. When it is a frequent allele, the chance that random subjects share an allele might increase somewhat. In a bad scenario (genotype error rate of 0.05, all errors produce the most frequent allele), the heterozygosity might drop so much that one additional marker (H≈0.8) is needed (data not shown).

Thus, genotyping errors often cause *in*consistent loci in related subjects, but only rarely increase the chance of consistent loci in *un*related subjects. To allow for a realistic rate of genotyping errors only two or three additional markers (heterozygosity ~0.8) are needed.

According to these theoretical considerations, it is possible to detect MZ twins and PO pairs with a limited number of polymorphic markers, while allowing for a reasonable number of genotyping errors. In the following, we will test this approach on a real data set.

## Materials

*Construction of the sample*

The PREVEND study is a large (N= 8 592) population based cohort study, conducted in the town of Groningen, the Netherlands [5]. The study was designed to investigate the etiology of microalbuminuria in relation to atherosclerosis. For this paper, two sub samples were used, that were genotyped for a genetic association study: firstly, a random subset; secondly, a subset of subjects selected on the prerequisite that some criteria were met that increase the likelihood of having family within the sample. The following criteria were used to select possibly related subjects: subject pairs that shared surname and address, and differ at most 12 years in age; or subject pairs that shared surname, and differ 18 to 40 years in age. Only subjects with surnames that are rare (<= 15 occurrences in the database) were considered, to increase the *a priori* likelihood of relatedness. This resulted in 689 randomly drawn subjects, and 619 subjects that were possibly related.

*Genotyping*
All 1308 selected subjects were genotyped for 28 microsatellite markers and 1 insertion/deletion polymorphism. The markers were in 11 genomic regions, at nine different chromosomes. Eighty subjects were discarded for further analysis, as they had less then six genotypes scored after one round of genotyping. We verified, using analysis of variance, that there were systematic differences in genotyping success between subjects, so this elimination is mainly caused by accidental lack of quality in DNA isolation [6]. After two rounds of genotyping, 97.8% of the alleles were scored for the remaining subjects, 638 random and 590 possibly related.

## Methods

*MZ twins*
The software GRR (graphical representation of relationship errors)[7] was used to obtain a visual representation of the genetic concordance within the data set, using average number and standard deviation of alleles *identical-by-state* (IBS). In the absence of marker mutations and genotyping errors, MZ twin pairs have IBS=2, standard deviation in IBS = 0, and are therefore expected in the lower right corner of the graph (fig. 2).

*PO pairs*
GRR can be used to visually detect MZ twins, but PO pairs are not readily identified. We created computer software (using Delphi 5 for Windows) to compare all possible subject-pairs and count for each pair the numbers of consistent, inconsistent and uninformative loci. All subject pairs fulfilling the criteria are saved to file. A permutation procedure was built-in to empirically estimate the chance of false-positive results. In the permutation procedure, the observed genotypes were permuted independently for each locus over all subjects (assuming linkage equilibrium between loci). As missing genotypes are not randomly distributed over the subjects [6], these are not permuted.

*Constraints on potential relationships*
Non-genetic information constrains inferred relationships. True MZ twins should share gender and age. For potential PO pairs the most important parameter is the age difference between both subjects: PO pairs typically differ between 18 and 40 years in age. When subjects are part of multiple pairs, additional rules are available: an offspring cannot have more than two parents, parents should be male and female, and both parents should contribute one allele to the offspring at all loci.

## Results

*MZ pairs.* Details are in **appendix 2**. Four MZ twin pairs were discovered, with high sensitivity and specificity. Total genotyping error rate is estimated 1.8%.

*PO pairs.* Details are in **appendix 2.** Starting with 47 putative PO pairs, 38 PO pairs, 8 sib pairs and one false positive pair were identified. Three trios and three sib pairs with one parent could be formed. The number of identified PO pairs was 4.1 times higher in the sub sample with possibly related subjects. Detectable genotyping error rate estimated in PO pairs was 0.0075 (SD 0.0022) per genotype (as expected less than half of the rate in MZ twins).
*Sensitivity:* Details are in **appendix 2.** Adding the missing 2.2% of the genotypes might contribute about 8 PO pairs. About 3 PO pairs are expected to be missed due to genotyping errors alone. Marker informativeness appeared to sufficient as only one false positive was found, although the number of independently segregating loci was insufficient to distinguish between PO pairs and sib pairs.

## Discussion

Samples drawn from a limited population have a certain chance of containing related individuals. It can be of interest to detect hidden relatedness, as increased relatedness increases the variance, especially when the relatives are asymmetrically distributed. Pedigree information can also be helpful to establish genetic phase, which can increase the amount of mapping information present in the data. In this paper, we present a rather intuitive procedure to detect parent-offspring and monozygotic twin pairs: using microsatellite markers genotypes (for example from candidate gene association studies), all subject pairs are considered in turn, to check whether it could form a potential parent-offspring or monozygotic twin pair. These relationships are genetically deterministic as long as marker mutations and genotyping errors can be excluded. Other types of relationships within pedigrees (siblings, grandchildren, cousins) are less deterministic, and thus more markers are needed to detect these.

When applying the procedure to a realistic data set, a number of complicating factors are met, most of which can be accounted for.

- *How many markers are needed to minimize chance concordance between unrelated subjects?* As large numbers of subject pairs are tested, the number of polymorphic markers needed to exclude chance results can be substantial. The number of markers needed to reach a desired confidence level can be estimated when sample size is known, on assumption of linkage equilibrium and assuming allele frequencies (table

1 and figure 1). When using highly polymorphic markers (heterozygosity 0.9), 13 loci might be sufficient for a sample of ~1400 subjects. However, when using markers as used in candidate gene association studies, heterozygosities are usually much lower. When markers with a heterozygosity of 0.7 are used, 43 independent loci are needed. When SNPs are used with a heterozygosity of 0.2 about 680 loci are needed. Sample size has limited influence: decreasing sample size from 1400 to 140 decreases the estimated number of markers required with only a third.
Linkage disequilibrium between the markers will increase the number of markers needed, relative to these estimates. Given the complex nature of the phenomenon, we cannot account for this in our estimation. Especially fine-mapping studies will suffer from this loss of power.
The numbers of markers needed might seem quite substantial with regard to most studies in the past. However, in practice multiple association studies involving different candidate genes are successively performed on the same population sample; combining the information of multiple studies might well yield the number of markers that is needed to perform these analyses. Secondly, highly polymorphic markers have relatively large influence on the discrimination. It's worth considering adding one or more highly polymorphic markers, only for this goal. Thirdly, whole genome SNP screens using thousands of markers are starting to be used [8, 9, 10]. Datasets of this size will undoubtedly carry enough information to permit the proposed procedures. The number of independently segregating loci will, of course, be much lower than the number of markers. This complicates analytical estimation of likelihoods, but will not invalidate the procedure itself.

- *Missing data.* When considering a pair of subjects, only loci that are fully known in both subjects are informative. In our data set with 97.8% of data known, only 50% of the subject pairs where completely informative, while 86% of the subject pairs had at most 2 missing loci. Thus, missing data have considerable influence on the *effective* number of markers that are available.
- *Genotyping errors* will often make consistent loci appear inconsistent. The opposite (inconsistent loci that appear consistent) occurs seldom. Due to the asymmetrical effect, the number of markers needed (to reliably distinguish related from unrelated subjects) will increase somewhat, but exact numbers depend on what type of error is made.
- *The presence in the data set of subject pairs that are otherwise related. Subjects* within a pedigree show much larger genetic concordance than unrelated subjects. Consequently, it is easier to distinguish parent-offspring pairs from unrelated subjects, than from other types of relationship. In our example data set, some of the subject pairs that were found genetically consistent with being parent-offspring pairs should be labeled sib-pairs. However, non-genetic information can be used to discriminate between PO pairs and sib-pairs with reasonable confidence: sib pairs differ mostly <12 years in age, while parent-offspring pairs differ at least 18 year. In

our data set no putative parent-offspring pairs were identified with an age difference 12-18 years (which is consistent with both types of relationships).
It would be of some interest to check the relatedness of identified subject pairs by interviewing the subjects. However, non-paternity occurs at a non-negligible but unknown rate within trios, and is often not admitted or unknown [11, 12]. This prevents information provided by the subjects to be used as gold standard. To respect the privacy of the subjects, we did not pursue this issue.

Although the detected number of parent-offspring pairs identified seems limited, crude estimations (appendix 2) suggest it might be increased compared to the general population. Another sub sample was enriched for containing parent-offspring pairs using surname - a heritable, non-genetic trait. Within this sub sample, the number of detected parent-offspring pairs was 4.1 times higher. This suggests that the number of relatives might be substantially increased in case-controls studies on hereditary traits, especially when a strongly genetic trait is studied in a smaller, more isolated population. This could potentially lead to smaller or larger confounding of the results.
We developed procedures to detect parent-offspring pairs and monozygotic twin pairs that are present genetic studies, and provide estimates of the numbers of markers needed to do so. Our method detected a small number of parent-offspring pairs hidden in a random population sample and a substantially higher number in a sample enriched for relatives. Routine genetic data can be used to detect first-degree related subjects, using a reasonable number of markers, given that these are highly polymorphic and reliable.

**Appendix 1: the number of markers needed, details.**

In this section algorithms are developed to compute an estimate of the numbers of markers needed to reliably detect parent-offspring pairs.
All possible two genotype combinations are enumerated; the probability of each genotype combination is computed as the product of the allele frequencies for marker *i*. The probabilities of genotype combinations that share at least one allele are summed. Call this probability $p_{sh(i)}$ (the chance that 2 random genotypes share at least one allele at marker *i*). Call $\alpha$ the chance of labeling a pair of unrelated subjects as related (i.e. a false-positive result).
Assuming that all *i* markers have the same $p_{sh}$ and are independent (i.e. when there is no linkage-disequilibrium (LD) between the markers; the subjects are unrelated), the probability of sharing at least one allele at *L* loci can be simplified: $\alpha = p_{sh}^{L}$. This can easily be rewritten to make the number of markers, *L*, a function of $p_{sh}$ and $\alpha$.

$$^{p_{sh}}\log(p_{sh}^{\ L}) = {}^{p_{sh}}\log(\alpha)$$

$$L \cdot 1 = {}^{p_{sh}}\log(\alpha)$$

When *L* is known for a certain level of $\alpha$, $L'$ can be easily deduced for another level of false-positive results $\alpha'$: *L* must be multiplied by $\log(\alpha')/\log(\alpha)$. For example, to compute the number of markers needed for $\alpha'=10^{-3}$ instead of $\alpha=10^{-6}$, *L* has to be multiplied by 3/6= 0.5.
To get an impression of L, it was computed for a range of simulated allele frequencies, representing markers with 2 to 11 alleles, randomly drawn from a uniform distribution. For each marker were calculated: heterozygosity, $p_{sh}$, and *L* when setting $\alpha$ to $10^{-6}$. Results are shown in **figure 1**. Average values for sets of twenty simulated markers with heterozygosities of 0.5, 0.6, 0.7, 0.8 and 0.9 are shown in **table 1.** In an analogous way, the same was done for monozygotic twins.

To support the notion that highly polymorphic markers have relatively large influence on the needed number of markers: consider the likelihood of two unrelated subjects sharing at least 1 allele at two independent markers. When both markers have heterozygosity 0.7 this has a probability $0.726 \cdot 0.726 \approx 0.53$, while heterozygosities of 0.5 and 0.9 (same average heterozygosity) result in a much lower probability, $0.353 \cdot 0.872 \approx 0.31$.

## Appendix 2 Details on identification of related subjects

*MZ pairs*
**Figure 2** shows the agreement between all subject pairs. MZ twins are expected to lie in the lower right corner of the graph. The graph shows four subject-pairs in that region. *Sensitivity:* The amount of separation between these pairs and the rest of the subject-pairs suggests that all MZ twins were identified.
*Reliability:* Three of the four putative MZ pairs consist of subjects of the same age and gender (as expected for MZ twins), but in one pair the subjects differ 11 years in recorded age. This amount of genetic concordance is very unlikely to occur within regular sib pairs or PO pairs (data not shown). Therefore, other explanations remain, such as: swapped or wrongly labeled DNA samples or an error that occurred in the descriptive data set.

*Genotyping error rate*
Assuming marker mutations to be absent, MZ twins can be used to estimate the genotyping error rate. When considered as three true MZ twins, there are inconsistencies on three loci out of 164 informative loci ≈ 1.8% genotyping error rate. When all four pairs are indeed MZ twins, there are four inconsistent loci in 218 informative genotypes ≈ 1.8% genotyping error rate.

*PO pairs*
**Table 3** summarizes the number of subject pairs that show the most consistent loci. Note that the permutation procedure assumes linkage equilibrium between the loci. The shaded lines mark the categories that clearly have more observations than can be explained by chance. Clearly, excess of subject pairs is observed in the categories with 28 or 29 consistent loci, and at most 1 uninformative locus. Subject pairs in these categories are therefore considered putative PO pairs. Other categories have more inconsistent loci that appear to exclude PO pair relatedness, and the observed number of pairs in these categories does not exceed chance.
**Table 4** summarizes the 47 putative PO pairs. As stated above, additional data put constraints on identified relationships. First, MZ twin pairs should be left out (N=2). Second, when regarding subjects that involved multiple subject pairs, three trio's (two parents, one offspring) and three triads (one parent, two offspring) could be formed. This implied two putative PO pairs to be sib pairs. Also, a new PO-pair was discovered (not included due to 2 inconsistent loci), and one putative PO-pair appeared to be unrelated (third 'parent' in a trio).
Third, when regarding age difference (which for PO pairs is expected to be between 18 and 40), another 5 putative sib pairs were identified, all with age difference <12 years.

Combining this information, 38 PO pairs were identified, 8 sib pairs and 4 MZ twins. Only 1 putative PO pair is a false positive result. Most sib pairs were found, when allowing one genotyping error.

Subjects originated from two sub samples. 14 PO pairs were found in the sub sample of possibly related subjects (N=590), 4 were found in the sub sample of randomly drawn subjects (N=638). When corrected for the number of pairs considered, PO pairs were discovered 4.1 times more often within the sample of possibly related subjects (95% confidence interval: 1.5 to 8.6, $p< 0.05$). This suggests that sampling of subjects partially and/or indirectly based on a hereditary trait could substantially increase the number of related subjects included in a sample.

**Table 3: Subject pairs with many loci consistent with being a PO pair**

| Category | | | Observation | Permutations | | |
|---|---|---|---|---|---|---|
| Inconsistent | Consistent | Uninformative | Number of subject pairs | Average number of pairs | 99th percentile of number of pairs | Putative PO pairs? |
| 0 | 29 | 0 | 19 | 0.052 | 1 | Yes |
| | 28 | 1 | 9 | 0.025 | 1 | Yes |
| | 27 | 2 | 2 | 0.032 | 1 | Yes |
| | 26 | 3 | 1 | 0.026 | 1 | No |
| | 25 | 4 | 0 | 0.018 | 1 | No |
| 1 | 28 | 0 | 17 | 1.06 | 4 | Maybe |
| | 27 | 1 | 4 | 0.761 | 4 | No |
| | 26 | 2 | 4 | 0.692 | 3 | No |
| | 25 | 3 | 2 | 0.592 | 3 | No |
| 2 | 27 | 0 | 22 | 12.438 | 21 | No |
| | 26 | 1 | 8 | 8.717 | 17 | No |
| | 25 | 2 | 5 | 7.188 | 15 | No |

The first three columns define a level of agreement between a pair of subjects (respectively, the number of inconsistent, consistent and uninformative loci). The following column shows the observed number of subject pairs that fall in that category. The next two columns show the results of 1000 permutations of the observed genotypes: respectively the average number of subject pairs in that category; and (upper) 99th percentile of the number of subject pairs in that category. Shaded lines mark categories with an observed number of pairs that clearly exceeds random expectation, and thus are putative PO pairs.

**Table 4: identified related subjects**

| Category | Initial | Correction using additional data | | | | Result | | |
|---|---|---|---|---|---|---|---|---|
| *consistent / inconsistent / uninformative loci* | *Putative PO* | *MZ* | *Sib, ped* | *Sib, age* | *False* | *PO* | *Sib* | *MZ* |
| 29 / 0 / 0 | 19 | 1 | | | | 18 | | 1 |
| 28 / 0 / 1 | 9 | | | 2 | | 7 | 2 | |
| 27 / 0 / 2 | 2 | 1 | | | | 1 | | 1 |
| 28 / 1 / 0 | 17 | | 2 | 3 | 1 | 11 | 5 | |
| Other | | 2 | 1 | | | 1* | 1* | 2 |
| **Total** | 47 | | | | | 38 | 8 | 4 |

Some of the putative PO pair turned out to be MZ twins, sib pairs (either based on identified pedigrees (ped) or age difference (age)) or false. Last three columns show the final numbers of identified subject pairs. * Implied by other relationships. The PO pair has 27 consistent, and 2 inconsistent loci.

**Detectable genotyping errors**

Within the identified 3 trio's the detected genotyping error rate is 2 errors in 261 genotypes= 0.0077 (±0.0054) per genotype. When using the fully informative PO pairs (including trio's and accounting for subjects that are involved in multiple pairs) the error rate is 12 errors in 1595 genotypes= 0.0075 (±0.0022) per genotype

*Sensitivity*

In this section, it will be tried to estimate the number of PO pairs that remain unidentified.

*a) Missing genotypes:* The observed numbers of missing genotypes are shown in **Table 5.** Over 50.1% of the subject-pairs could be compared at all loci, while 84% of the pairs could be compared at at least 27 loci. The results mentioned above suggest that parent-offspring pairs with up to two uninformative loci can be identified with reasonable certainty. 84% of the subject pairs fall in this category, so further genotyping is expected to add about (1-0.84)/0.84 = 19%, i.e. about five PO pairs without genotyping errors, and three PO pairs with one genotyping error.

*b) Marker informativeness:* Eight sib pairs were identified that fitted the genetic criteria for PO pairs, although most with one inconsistent locus. This indicates that the marker informativeness is not sufficient to reliably distinguish between parent-offspring pairs and sib pairs. However, the distinction between first-degree related subjects and unrelated subjects can be made with considerable confidence, as indicated by the permutations and the fact that only one identified subject-pair is probably unrelated.

*c) Genotyping errors.* The frequency of errors that lead to an inconsistent locus (i.e. detectable errors) was estimated above at ~0.0075. Using this figure, 64.6% of the fully typed parent-offspring pairs are estimated to contain no errors, and 28.3%

**Table 5: amount of informative loci**

| Uninformative loci | Number of subject pairs | Percentage |
|---|---|---|
| 0 | 377146 | 50.1% |
| 1 | 164910 | 21.9% |
| 2 | 90259 | 12.0% |
| 3 | 51357 | 6.8% |
| 4 | 27336 | 3.6% |
| ≥5 | 42370 | 5.6% |
| Total | 753378 | 100% |

exactly one error. Taken together, 92.9% of the fully typed PO pairs will have at most one genotyping error, 26 pairs were observed in these categories. Therefore, of the fully typed pairs (50.1% of the sample) three PO pairs may be expected to have more than one genotyping error.

How many PO pairs are expected in Groningen? We can only provide crude theoretical estimates: in Groningen live about 90 000 subjects aged between 28 and 75 [1]. Suppose this population to consist entirely of 30 000 trios (2 parents, 1 offspring). Then 60 000 PO pairs are present in $(N \cdot (N-1)/2 \approx) 4.05 \cdot 10^9$ subject pairs. Frequency is then $1.5 \cdot 10^{-5}$.
Suppose each subject to be part in exactly one PO pair. Then 45 000 PO pairs are present. Frequency is then $1.1 \cdot 10^{-5}$. Actual frequency will be lower due to an unknown amount of migration. Increasing pedigree size will increase the frequency. We observed 4 PO pairs in 638 random subjects; frequency = $2.0 \cdot 10^{-5}$. This suggests that the observed number of PO pairs might be increased, due to preferential sampling of relatives.

**Software**
Source code of the software is available from the authors on request.
GRR can be obtained at: www.sph.umich.edu/csg/abecasis/grr/software.html

**References:**
[1] Statline.cbs.nl, Statistics Netherlands
[2] Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibly RM. Likelihood-based estimation of microsatellite mutation rates. Genetics. 2003; 164: 781-7.
[3] Gordon D, Heath SC, Ott J. True Pedigree Errors More Frequent Than Apparent Errors for Single Nucleotide Polymorphisms. Hum Hered 1999; 49: 65–70.
[4] Douglas JA, Skol AD, Boehnke M. Probability of Detection of Genotyping Errors and Mutations as Inheritance Inconsistencies in Nuclear-Family Data. Am. J. Hum. Genet. 2002; 70: 487–95.
[5] Pinto-Sietsma SJ, Janssen WM, Hillege HL, Navis G, De Zeeuw D, de Jong PE. Urinary albumin excretion is associated with renal functional abnormalities in a nondiabetic population. J Am Soc Nephrol 2000; 11: 1882–8.

[6] Spijker GT, Bruinenberg M, Te Meerman GJ. Efficiency control in large-scale genotyping using analysis of variance. Appl Biochem Biotechnol. 2005; 120: 29-36.
[7] Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of relationship errors. Bioinformatics. 2001; 17: 742-3.
[8] Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet. 2006; 38:617-9.
[9] Yamazaki K, McGovern D, Ragoussis J, Paolucci M, Butler H, Jewell D, Cardon L, Takazoe M, Tanaka T, Ichimori T, Saito S, Sekine A, Iida A, Takahashi A, Tsunoda T, Lathrop M, Nakamura Y. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet. 2005; 14: 3499-506.
[10] Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002; 32: 650-4.
[11] Macintyre S, Sooman A. Non-paternity and prenatal genetic screening. Lancet 1991; 338: 869-71.
[12] Le Roux MG, Pascal O, Andre MT, Herbert O, David A, Moisan JP. Non-paternity and genetic counseling. Lancet 1992; 340: 608.